DOCUMENT RESUME

ED 274 706 TM 860 583

AUTHOR	Ackerman, Terry A.
TITLE	Use of the Graded Response IRT Model to Assess the Reliability of Direct and Indirect Measures of Writing Assessment.
PUB DATE	Apr 86
NOTE	43p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Correlation; *Essay Tests; High Schools; Interrater Reliability; *Latent Trait Theory; Measurement Techniques; Research Methodology; *Test Reliability; Test Theory; *Writing Evaluation; *Writing Skills
IDENTIFIERS	*Graded Response Model

ABSTRACT

The purpose of this paper is to compare the precision of direct and indirect measures of writing assessment using the test information functions from a graded response Item Response Theory (IRT) model. Subjects were 192 sophomore English students from a parochial high school in Wisconsin. Both direct and indirect measures of writing ability were used. Comparisons between the IRT information functions for the three analytic raters in five different writing skill areas are also examined. Comparisons are also made with results obtained using classical test theory methodology. Results show that the plots of the IRT information functions can be used to provide valuable information about essay raters. However, correlation coefficients between the examinee IRT ability estimates calibrated from the three sets of ratings and the standardized test separately, were found to be quite small. The implications of these findings and directions for future research are discussed. (Author/JAZ)

Use of the Graded Response IRT Model to Assess the Reliability of Direct and Indirect Measures of Writing Assessment

Terry A. Ackerman

The American College Testing Program

Running Head: Graded Response Application to Writing Assessment

Abstract

The purpose of this paper is to compare the precision of direct and indirect measures of writing assessment using the test information functions from a graded response IRT model. Comparisons between the IRT information functions for the three analytic raters in five different writing skill areas are also examined. Comparisons are also made with results obtained using classical test theory methodology. Results show that the plots of the IRT information functions can be used to provide valuable information about essay raters. However, correlation coefficients between the examinee IRT ability estimates calibrated from the three sets of ratings and the standardized test separately, were found to be quite small. The implications of these findings and directions for future research are discussed.

Use of the Graded Response IRT Model to Assess the Reliability of Direct and Indirect Measures of Writing Assessment

One of the most serious concerns of direct writing assessment is the lack of a standard methodology for use in assessing the reliability of essay scoring. Essay ratings do not provide the convenience that multiple choice items do in computing reliability coefficients such as KR-20. One method from classical test theory, is to examine the ratio of the estimated variance components to obtain generalizability coefficients. Another relatively new approach which could be used to assess the accuracy of essay ratings is item response theory (IRT). It is the purpose of this study to demonstrate how the graded response IRT model can be effectively used to assess not only the precision of raters, but also provide comparisons between the information being acquired from direct and indirect measures of writing. Advantages of using item response theory are discussed.

Theoretical Background

Samejima (1969) deve'oped a graded response IRT model which can be used to estimate item difficulty and discrimination parameters and examinee ability parameters from rating scale data. For $m_g + 1$ categories or ratings, item scores defined for item g can be denoted as x_g. The possible values or ratings of x_g are successive integers, 0 to m_g. For each individual responding to n items (or receiving n ratings), there is a response pattern or vector of integers, $v = (x_1, \ldots x_n)$. A given response to an item may be obtained from respondents at different points on an ability continuum. The probability of any particular response to an item for a given ability (θ) is defined as the operating characteristic.

The probability of an individual responding in any particular category r_g for item g is given by the formula

$$P_{x_g} = P(r = x_g \,|\theta) = P^*(x_g - 1|\theta) - P^*(x_g|\theta)$$

where r = ordered responses 1, 2, 3 ... m_g, and

$$P^*(x_g - 1|\theta) = \left[1 + \exp\left\{-a_g(\theta_j - b_{x_g - 1})\right\}\right]^{-1},$$

$$P^*(x_g|\theta) = \left[1 + \exp\left\{-a_g(\theta_j - b_{x_g})\right\}\right]^{-1},$$

a_g is the discrimination parameter for item g,

b_{x_g} is the threshold or difficulty parameter for category m_g, and

θ_j is the ability level of person j.

It is further assumed that

$$P^*(0|\theta) = 1 \text{ and } P^*(m_g|\theta) = 0.$$

Since there are $(m_g + 1)$ categories of response, there are m_g cutting points or thresholds, for an item, as well as m_g expressions for P_{x_g}. The probability P_{x_g} is a monotonically increasing function of θ in the range $-\infty < \theta < \infty$ with a constant discrimination parameter. This implies that the difficulty value of a category, b_{x_g}, continually increases, with b_{m_g} denoting the greatest difficulty.

There is only one "a" parameter for each item because the model assumes that the discriminating power is constant throughout the whole thinking process required to solve a problem or to rate an examinee on a particular topic (Samejima, 1969, p. 19).

Reliability: Classical Test Theory vs. IRT

In classical test theory the reliability coefficient is oftentimes computed using a correlation coefficient which implies that its value depends not only on the test itself, but also upon the specific group of examinees that were tested. The generalizability of coefficient alpha and Kuder-Richardson formulas 20 and 21 is rather limited because they are defined within a specific group of examinees. The standard error of measurement in most cases is derived from the reliability coefficient, and, therefore, cannot be interpreted as inherent to a particular test.

Lord (1980, p. 52) presents the formula for an IRT counterpart to the classical test theory reliability coefficient which is based upon the estimated item parameters and abilities. These parameter estimates are independent of the group of examinees, and are therefore intrinsic to the test items. However, for the most part, in item response theory it is the test information function that is used to specify the measurement precision of a test.

Several advantages exist in using the IRT test information function. One important advantage is that the contribution of each item to the total test information is additive. That is, the contribution of each item is independent of the other items in the test; therefore, the effect and impact of an item on a test can be readily determined. In classical test theory, the contribution of each item to the total test reliability depends largely on how each item correlates with other items in the test. Specifically, the property of independent item contributions is not present (Gulliksen, 1950).

A second major advantage is that an information value can be obtained for each value on the ability scale. In classical test theory when the

reliability coefficient of a test is computed, the standard error of measurement is usually assumed to be constant throughout the entire ability range (although standard errors of specific test scores can be estimated separately.) However, in IRT, the information function is conditioned on θ, and a test can be determined to be more or less precise at different ability levels. By comparing the information functions of two items (or tests) one can determine the ability range in which one item (or test) is providing more information (measuring more precisely) than another.

Samejima (1969) derived the information function for a given graded item

$$I_g(\theta) = \sum_{x_g = 0}^{m_g} \frac{\{P^{*\prime}_{x_g} - P^{*}_{(x_g + 1)}\}^2}{\{P^{*}_{x_g} - P^{*}_{(x_g + 1)}\}}$$

where $P^{*}_{x_g} = [1 + \exp\{-a_g(\theta_j - b_{x_g})\}]$

and $P^{*\prime}_{x_g}$ is the first derivative of $P^{*}_{x_g}$ with respect to θ, which is

$$P^{*\prime}_{x_g} = \frac{a_g \exp\{-a_g(\theta_j - b_{x_g})\}}{[1 + \exp\{-a_g(\theta - b_{x_g})\}]^2} .$$

For each $m_g + 1$ category, an information function can be calculated. The item information function is thus the sum of the individual m_g graded category information functions. The information function for an entire test or series of ratings is simply the sum of the individual item information functions.

It is the purpose of this paper to calibrate samples of both direct and indirect writing assessment using the graded response IRT model. Using both the item parameter estimates and the information functions, comparisons between each of the three raters as well as comparisons between the raters and the standardized test will be examined. The strengths of using IRT to evaluate the precision of each type of writing assessment are discussed.

Method

Subjects

The subjects who participated in this study were sophomore English students from a parochial high school in southeastern Wisconsin. The subjects consisted of the entire sophomore class except for eleven students who were absent on days when the writing tests used in this study were administered. Eighty-five of the 192 subjects were female, 107 were male.

Materials

Both direct and indirect measures of writing ability were used in this study. The direct measure consisted of an expository essay with students having a 50-minute class period in which to write about the "beneficial aspects of television."

One week after the students had written the essays, they were administered an indirect measure of writing ability, the Language and Writing subtests of the Comprehensive Assessment Program. The standardized subtests used in this study measured five areas of writing skill: spelling (SP), capitalization-punctuation (CP), correct expression (CE), usage (US), and paragraph development (PD). In all, the subtests totalled 70 items, and required a total administration time of 75 minutes.

The standardized test was machine scored. Students' item responses were then recoded into two graded categories, either as 0 (incorrect) or 1 (correct). The essays were analytically scored by three trained English

teachers. Ratings were given in the same five areas of writing skill as the standardized subtests. The scoring consisted of counting the number of errors in each skill area, and then transforming the number of errors into rating categories. The spelling errors were transformed into five graded categories, the punctuation into six graded categories, and correct expression, usage and paragraph development were each transformed into four graded categories. The categories were selected so there were approximately an equal number of students per category.

It was assumed that the 5 skill areas were assessed independently and thus the IRT assumption of local independence was not violated. Raters were specifically instructed to avoid "halo effects". That is, raters were trained not to let a high or low rating in one skill area influence ratings in the other areas.

Procedure

Three separate sets of graded response IRT calibrations were conducted using the computer program, MULTILOG (Thissen, 1985). In the first calibration, the dichotomous response data from the standardized test were analyzed, and item parameter and examinee abilities were estimated. In the second analysis the essay ratings from each rater, the ratings averaged over all three raters, and the three sets of ratings combined were calibrated separately. Each rating was considered as one item, thus five items or skill ratings were calibrated for each rater analysis and 15 items in the combined analysis. Separate estimates of examinee abilities were also obtained in each of the five computer runs. In the final analysis, the standardized test (70 items), the individual essay ratings (5 ratings X 3 raters), and the average

of the essay ratings (5 ratings) were combined and analyzed. This enabled the direct and indirect writing item parameter estimates to be placed on the same scale for comparison purposes.

The results of the first two sets of calibration analyses were used to determine the degree of agreement between writing ability estimates obtained by each reader and the standardized test. The third analysis was used to compare the amount of information provided by each rater and by each type of assessment within each skill area.

Results

Descriptive statistics summarizing the results of the essay test are reported in Table 1. The average number of errors identified by each rater within each skill area are listed in columns 2, 5, and 8. The results of the ratings when averaged over all three raters are displayed in column 11. Correlations between the number of detected errors and the corrected total number of errors (i.e., with the number of errors for the particular skill area being deleted) are shown in columns 4, 7, 10 and 13. These correlations provide an indication of how well each rater discriminated within each skill area.

Insert Table 1 about here

The most errors were detected in the capitalization-punctuation skill area; the least in usage. The biggest difference between the number of errors

detected by each rater was in the capitalization-punctuation skill area. Both raters 1 and 2 detected almost three times as many capitalization-punctuation errors as rater 3. The smallest difference seems to appear in paragraph development, where the mean number of errors detected for raters 1, 2, and 3 was .93, .88, and 1.10, respectively. In all cases, except the paragraph development area for rater 3, the standard deviations are larger than the mean number of errors detected indicating the positive skewness of the distribution of detected errors.

The correlation coefficients indicating the degree of discrimination of each rater for each skill area are quite small ranging from .03 to .20 for rater 1, .05 to .25 for rater 2, and .01 to .27 for rater 3. The low correlations may be due to lack of "internal consistency" within the 5 skill area ratings. That is, when correlations between the skill areas for each rater were examined, they were found to also be near zero, with some even being negative. No identifiable pattern could be found in their scatterplots. (However, this does provide evidence that the assumption of local independence was not violated.)

Interrater correlations between the number of detected errors for each skill area are reported in Table 2. The three raters had the greatest amount of agreement in spelling with the largest correlation being between rater 2 and rater 3, $r = .73$. The smallest correlation was between Rater 1 and Rater 3, $r = -.02$, in the correct expression skill area. Overall, the correlations are quite low, indicating lack of agreement in detecting errors in the five skill areas, particularly in the areas of correct expression and usage.

Insert Table 2 about here

Results of the standardized test are summarized in Table 3. The most difficult subtest was the Usage subtest, which had a mean number correct of 9.76 items out of a possible 18 (54%). The second most difficult test was the capitalization-punctuation subtest in which students only answered on the average 6.88 items correctly out of a possible 12 (57%). The easiest subtest was the correct expression subtest, with the average number correct equalling 14.36 out of a possible 18 (80%).

Insert Table 3 about here

Comparison of Essay Raters: Ability Estimation

The correlations of the examinee ability estimates based upon the essay and the standardized test are shown in Table 4. Above the diagonal the coefficients represent the interrater correlations between the total raw scores (ratings). Below the diagonal are the coefficients representing the correlations between the graded response IRT ability estimates. IRT reliability coefficients using the item and ability parameter estimates (see Lord, 1980, p. 52) are reported along the diagonal.

The greatest amount of interrater agreement for the raw score totals was between Rater 1 and Rater 2, r = .714. However, the greatest amount of

agreement for the IRT ability estimates was between Raters 1 and 2, r = .353. Rater 2 had the highest correlation with the standardized test, r = .259. When the ratings were combined the correlation of the IRT abilities dropped to .163, which was lower than the relationship between Raters 1 and 2, and the standardized test. Scatterplots of the ability estimates for all possible pairings of raters were examined to help explain the low interrater ability correlations, however, no particular pattern (e.g., curvilinearity) could be discerned.

Insert Table 4 about here

For each of the raters, the average rating,and the ratings combined the reliability coefficients are quite low, ranging from .148 (Rater 3) to .288 (Avg.) The standardized test IRT reliability coefficient is considerably higher, .731. The low reliability coefficients suggest that the theta estimates are not very accurate.

Thus the lack of precision in the IRT ability estimation is probably due to several things including a small number of items (ratings) per individual, and inconsistent response patterns among the individuals (e.g. somewhat random high and low ratings throughout the five categories). Since the response patterns are so varied, it was thought that the essay rating data might not be unidimensional. However, separate principal component analyses of each of the three sets of ratings for each rater were found to each have yield one principal component, suggesting the ratings probably are unidimensional.

Comparison of Essay Raters: Parameter Estimation

Although measures of difficulty and discrimination were estimated from the essay ratings, they remain dependent upon the group of examinees used in this study. Using the graded response IRT, both difficulty and discrimination parameters were also calibrated for the essay ratings. However, these parameter estimates are intrinsic to the rater and are independent of any group of examinees.

The graded response IRT parameter estimates for the essay ratings are reported in Table 5. These estimates were calibrated along with the standardized test items and are therefore more stable than those used to obtain the ability estimates. The parameter estimates are reported for each of the five essay skill areas. The a-parameter estimates reported in the first column represent how well the raters discriminated between the examinees. The b-parameter estimates represent the degree of difficulty of each of the categories within the particular skill areas.

Insert Table 5 about here

Rater 2 was able to discriminate between individuals best for spelling, capitalization-punctuation and correct expression. Rater 3 was best at discriminating between individuals in usage, and Rater 1 was best for paragraph development. Overall, the raters were best at discriminating between individuals in spelling and capitalization and poorest in paragraph development. The rankings of the raters within each skill area based upon the

IRT discrimination estimates were identical (except for correct expression) to the rankings based upon the discrimination correlations reported in Table 1.

By examining the threshold (difficulty) parameter estimates, one is able to determine the strictness of the individual raters. For example, in spelling the difficulty parameters indicate much easier ratings by Rater 3. This would imply that, relative to Raters 1 and 2, Rater 3 was able to detect fewer spelling errors. The same pattern holds for the capitalization-punctuation category. Rater 2 is the most stringent in both the correct expression and usage categories. Rater 1 was most stringent in rating paragraph development. Since the threshold parameters are all on the same scale, one can also compare the strictness between skill areas. Specifically, by examining the b estimates it can be seen that the raters were able to detect more errors in capitalization-punctuation than any other category The smallest number of errors was detected in the usage category. These results, for the most part, coincide with the mean number of detected errors reported in Table 1.

Ideally one might expect that the raters could discriminate between individuals equally well in all skill areas. Likewise, the consistency in the different thresholds for each category within the skill area should be nearly the same across all the raters.

Comparison of Raters: Measurement Precision

The information function is directly related to the discrimination parameter, thus the more a rater can discriminate between two examinees, the more precise the measurement process. Figures 1 through 5 show the plots of the information function values for each rater for each category. By

examining these plots one can see which rater is most reliable and over what ability range. Unlike classical test theory which usually assumes the error of measurement is constant over the entire ability range, IRT models assume that raters or items are not equally informative at each ability level. For example, in Figure 1, it can be seen that Rater 2 is the most reliable in assessing spelling, but only in the ability range from -2.5 to 1.0. In the higher ability range ($\theta > 1.0$) Rater 1 is most informative and in the lower ability range ($\theta < -2.5$) Rater 3 is most precise.

Insert Figures 1-5 about here

When Figures 1 through 5 are considered in concert, it can be seen that the raters are most reliable in spelling and capitalization-punctuation. In comparison, the raters provided very little information in judging the essays on paragraph development.

In the ideal setting, one would hope that the rater's information curves would be both similar (representing a close agreement of the construct to be rated) and consistently high across the entire test population ability range (implying that the measurement process is highly reliable for all examinees).

Direct vs. Indirect Assessment

To provide a stable comparison, between a single rater and the standardized test, the essay ratings were averaged over the three raters. The estimated parameters for the average rating are reported in Table 5. The

corresponding standardized test item parameter estimates are displayed by category in Table 6.

Insert Table 6 about here

For the most part, the items on the standardized test were quite easy as evidenced by the negative difficulty parameter estimates. The most difficult subtest was the usage subtest, in which only 38% of the b estimates were greater than zero. There is also a preponderance of low discrimination estimates. Eleven of the items have a estimates less than .3. Items with low $\hat{a}$ values have essentially flat ICC's and thus provide very little information.

Figures 6-10 show the plots of the item informations curves along with the total for each of the standardized subtests. Notice that the scale along the y-axis has been changed from Figures 1-5 since more information is being provided for some skill areas.

Insert Figures 6-10 about here

The most information provided by any one subtest, spelling, was 3.51. Ironically, this subtest has the fewest number of items, only ten. The least amount of information is provided by the capitalization-punctuation subtest which has 12 items. Most of the information being provided by the subtests is

over the low ability range $-3 < \theta < 0$. The most information provided over the upper ability range, $1 \leq \theta \leq 3$, is provided by 18 standardized usage questions.

The information curves for the average essay ratings are shown in Figures 11-15. The total information function for each skill area rating, represented as a chain dot curve, is the sum of the information curves for the graded categories. The information provided by each of the graded categories are represented by solid curves. Notice also that the scale along the y-axis has been changed from that used for the standardized test so that the relatively small amount of information provided by some of the categories can be noticeably represented.

Insert Figures 11-15 about here

As in the standardized test, the spelling ratings were most informative, followed by the capitalization-punctuation ratings. The paragraph development and the usage ratings provided the least amount of information.

The information curves for the categories within each skill area, should be about the same height and equally spaced, which would mean that each of the categories are being assessed with the same degree of precision over equal intervals of ability. For example, the level information curves within the capitalization-punctuation ratings show more variance among the six category information curves than do the spelling category curves. Ideally, the "total" information curve should be equally high across the targeted ability range for each of the essay skill areas.

To provide an overall graphic comparison between the standardized test and the average essay ratings, the test information curve for both the essay and the standardized test were plotted. The results are shown in Figure 16.

Insert Figure 16 about here

As shown in Figure 16, the standardized test provides more information throughout the entire ability range, particularly in the abilities range below -1.0. The maximum information value of the standardized test is 8.56 at $\theta = -2.0$; for the essay, the maximum information is 4.2 at $\theta = -1.4$.

As a final analysis, plots showing the relative efficiency of the essay to the standardized test were drawn. Relative efficiency (RE) of a test score y with respect to a test score x is simply the ratio of their information functions:

$$RE\ \{y,x\} = \frac{I\ (\theta,y)}{I\ (\theta,x)}\ ,$$

when θ in $I(\theta,y)$ and $I(\theta,x)$ are the same. (It's important to note that the information function itself cannot be interpreted in an absolute sense unless a valid θ metric is defined. However, the relative efficiency ratio is invariant under any monotonic transformation of the ability scale.)

The relative efficiency ratio is an example of another advantage IRT theory has over cla test theory. In classical test theory the reliability of t nnot be compared directly unless given to the exact same examinees. Ra comparing t reliability coefficients, the

Spearman Brown prophecy formula is usually invoked to indicate how many more parallel items would have to be added to each test to reach a particular reliability standard (i.e. r = .90).

By using the IRT model's relative efficiency ratio, not only can the precision of measurement between two tests be compared, but differences can be computed at each point along the ability range. Thus one can examine a relative efficiency plot and not only identify which test is measuring more precisely, but also how many items and what types (i.e. difficult, easy) of items need to be added to make the precision of measurement the same in both tests. More importantly, the tests do not have to be administered to the same examinees.

In Figure 17, the relative efficiency for the total of the average essay ratings compared to the total standardized test is shown. The solid horizontal line at RE = 1.0, represents the relative efficiency if the information provided by the essay ratings were equal to the amount of information provided by the standardized test at all levels of ability. The dotted horizontal line at RE = .07 represents the ratio of number of "items" in the essay test divided by the number of items in the standardized test. Assuming each item on both measures provided exactly the same amount of information, this dotted line would be the expected relative efficiency if the two tests differed only in length.

In the θ range from -3 to -2, it can be seen that the curve representing the relative efficiency of the essay test to the standardized test has a value of about .25, which means it is about 25 percent as informative as the standardized test; thus four times as many ratings would have to be added to make the tests equally accurate in these ability ranges. In the ability range from $0 < \theta < 1.0$, the number of essay ratings would have to be approximately doubled to be as effective.

Insert Figure 17 about here

The plot of the essay relative efficiency curve in Figure 17 is somewhat misleading since it might be assumed that the standardized test is the better method of assessment throughout each of the five skill areas. However, the plots of the relative efficiency curves by skill areas (e.g., $I_{\text{Essay SP}}/I_{\text{Standardized SP}}$), shown in Figure 18, clearly demonstrate that the average essay rating for capitalization-punctuation and spelling provided more information than its standardized counterpart. In fact, at $\theta = .95$, the capitalization-punctuation rating is over twice as informative as its standardized counterpart.

Insert Figure 18 about here

Discussion

One of the shortcomings of this study is the small number of skill area ratings per examinee. That is, each essay rater only provided five scores per examinee. This appears to be too few "items" to provide accurate IRT ability estimates as evidenced in the low interrater correlations in Table 3. Rather than increasing the number of skill areas assessed per essay, a better approach might be to have the students write several essays. Likewise, it

suggests that before a classroom teacher can precisely assess a student's writing ability, several pieces of writing should be obtained.

Another weakness of the study was the low interrater correlations between detected errors within each skill area. The pattern of correlations (Table 2) appears to parallel the difficulty in defining the skill area. That is, the greatest amount of agreement was in spelling, an area which is well defined; the least amount of agreement was in correct expression, an area where grading may be subject to individual interpretation.

One of the strengths offered by using the graded response information function is that it provides a means by which raters' precision can be assessed over an entire ability range. Figures 1-5 provide a good example of how raters can be compared using this approach. Several things can be easily discerned from these plots including the ability range in which each rater is most precise, which rater is overall most precise and where, and the similarity of the raters. These comparisons not only could be used in training raters, but could be used as a monitoring process to insure the stability and accuracy of the raters. Although not very applicable to the classroom teacher, such analyses would seem to be valuable in a large scale essay test such as the Test of Standard Written English (ETS) or Advanced Placement Examinations (College Board).

Test information functions could also be quite useful in evaluating the quality of items on a standardized test. Not only could poor items which provide little or no information be weeded out, but one could check to see if the overall test is providing adequate information in the targeted ability range.

Although it might be argued that comparing the results of just a single set of essay ratings with the results from a standardized test is not valid,

it is interesting to note that the average essay ratering provided more information in the spelling and capitalization-punctuation skill areas for certain ability ranges than the standardized test! A possible explanation for this result, is that in these skill areas each word on an essay could be considered to be an "item". Thus, when considered from this perspective, these essay skill areas have many more items than their standardized counterpart and might be expected to be more informative. Notice also that these two skill areas had more graded categories than the other skill areas.

Conclusion

The results of this study demonstrate that the graded response IRT model and its corresponding information functions provide the methodology necessary to make valuable comparisons between essay raters and indirect measures of writing. Plots of the information functions provided knowledge about the accuracy of the raters within each skill area. However, the IRT ability estimates calibrated on the ratings of a single rater were found to have low reliability and did not correlate with those from other raters, nor with those from the standardized test. This, in part, may be due to the small number of "items" (5 ratings) and the small sample size. Although when the ratings from the three raters were combined (i.e., 15 ratings) the correlation with the estimated IRT abililties from the standardized test dropped.

Discrimination and difficulty parameter estimates calculated using both IRT and classical test theory were found to be in close agreement suggesting that one may not have to go through expensive IRT calibration to rank raters on these parameters.

These results suggest several possible directions for future research. The lack of precision in the graded response IRT ability estimates using essay ratings needs to be further explored. One might suspect that by increasing the number of essays used in the calibration process, the ability estimates would become more precise. The lack relationship between the essay and standardized test ability estimates also needs to be examined. Although some may argue this issue is more a question of validity than reliability.

Another avenue of future research, if the use of the graded response IRT calibration process was facilitated, would be to compare different types of essays. That is, perhaps certain types of essays (e.g. expository, narrative) might prove to be more informative at different ability levels. Such knowledge might be useful in planning the sequence of writing instruction.

References

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer Nijhoff Publishing.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, 17.

Samejima, F. (1977). A use of the information function in tailored testing. Applied Psychological Measurement, 1, 233-247.

Table 1
Means, standard deviations and correlations of the number of detected errors in Each Skill Area

Skill Area	Rater 1			Rater 2			Rater 3			Average		
	$\bar{X}$	s	r^a	$\bar{X}$	s	r^a	$\bar{X}$	s	r^a	$\bar{X}$	s	r^a
SP	1.86	2.21	.12	1.33	1.91	.21	.80	1.61	.15	1.33	1.92	.22
CP	3.20	3.38	.03	2.35	2.81	.25	1.10	1.72	.24	2.22	2.72	.23
CE	.38	.69	.20	.73	1.07	.16	.95	1.00	.01	.69	.93	.23
US	.39	.71	.04	.49	.89	.18	.09	.36	.27	.32	.69	.24
PD	.93	1.12	.14	.88	.99	.05	1.10	.96	.02	.97	1.02	.24

Note: [a]Correlations between the number of detected errors and the corrected total number of detected errors (i.e., with the number of detected errors for the skill area in question deleted.)

Table 2

Interrater correlations between the number of errors detected within each writing skill area

Subscore	Interrater Correlations		
	r_{12} [a]	r_{23} [a]	r_{23} [c]
SP	.64	.55	.73
CP	.53	.55	.49
CE	.19	-.02	.17
US	.25	.14	.20
PD	.33	.39	.31
Total	.67	.65	.65

Note: [a]Correlation between Rater 1 and Rater 2.

[a]Correlation between Rater 2 and Rater 3.

[a]Correlation between Rater 2 and Rater 3.

Table 3
Standardized Test Results

Skill Area	n[a]	$\bar{X}$	σ
SP	10	7.78	1.69
CP	12	6.88	1.87
CE	18	14.36	2.11
US	18	9.76	2.48
PD	12	9.15	2.04

Note: [a]The number of items in the designated subtest.

Table 4
Correlations Between the Ability Estimates Derived from the Three Essay Raters and the Standardized Test.

	Rater 1	Rater 2	Rater 3	Avg[a]	Comb[b]	Stand[c]
Rater 1	.164[d]	.714	.524	.850	.882	.456
Rater 2	.282	.276[d]	.553	.848	.900	.493
Rater 3	.185	.353	.148[d]	.760	.780	.257
Avg	.418	.359	.479	.288[d]	.958	.456
Comb	.198	.646	.804	.512	.228[d]	.481
Stand	.259	.237	.141	.233	.163	.731[d]

Note: The coefficients above the diagonal represent the interrater correlations between the total raw scores (ratings). The coefficients below the diagonal represent the interrator correlations between the IRT ability estimates.

[a] Ability estimates based upon the ratings of the three raters averaged

[b] Ability estimates based upon the ratings of the three raters combined.

[c] Ability estimates based upon the standardized test.

[d] IRT reliability coefficients (see Lord, 1980, p. 52).

Table 5
Essay Parameter Estimates Arranged by Rater Within Skill Area

Rater	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{b}_5$
			Spelling			
1	1.31	-1.81	-1.09	-0.23	0.96	
2	1.74	-1.73	-1.32	-0.67	0.17	
3	1.32	-3.24	-2.25	-1.40	-0.38	
AVG	2.04	-2.13	-1.45	-0.90	-0.08	
			Capitalization/Punctuation			
1	1.22	-1.63	-0.82	-0.17	0.53	1.47
2	1.67	-1.69	-0.91	-0.55	-0.04	0.63
3	1.37	-3.17	-2.00	-1.70	-1.02	-0.13
AVG	1.84	-1.75	-1.26	-0.81	-0.34	0.66
			Correct Expression			
1	0.56	-7.45	-4.30	-1.82		
2	0.91	-3.10	-1.69	-0.47		
3	0.26	-8.91	-4.37	1.56		
AVG	1.03	-5.54	-3.31	-0.69		
			Usage			
1	0.41	-9.08	-6.22	-2.25		
2	0.76	-5.36	-3.08	-1.01		
3	1.18	-8.05	-3.71	-2.56		
AVG	0.95	-9.19	-4.73	-2.19		

Table 5 (cont.)

Rater	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{b}_5$
			Paragraph Development			
1	0.79	-2.91	-1.67	0.27		
2	0.32	-8.48	-4.20	1.05		
3	0.37	-7.65	-2.86	2.84		
AVG	0.77	-5.30	-3.06	0.08		

Table 6
Standardized Test Item Parameter Estimates Arranged by Skill Area

Item	$\hat{a}$	$\hat{b}$	Item	$\hat{a}$	$\hat{b}$
			Spelling		
1	.90	-3.84	6	.76	-.50
2	1.66	-2.23	7	.52	1.05
3	.50	-7.96	8	1.03	-.58
4	1.45	-1.97	9	1.36	-2.04
5	2.34	-2.06	10	.63	-.87
			Capitalization/Punctuation		
1	.15	-2.40	7	.69	-2.13
2	.90	-1.64	8	.56	1.33
3	.53	-1.95	9	.11	-1.32
4	.92	-1.97	10	.38	2.66
5	.79	-1.54	11	.02	14.91
6	.10	7.63	12	.20	-.46
			Correct Expression		
1	1.34	-2.93	10	.23	-.14
2	.97	-2.26	11	4.17	-6.08
3	.10	-11.55	12	1.13	-1.82
4	.57	-2.89	13	.29	-.18
5	.52	-1.64	14	.71	-3.63
6	.11	-24.79	15	.63	-0.11
7	1.06	-3.83	16	.56	-3.26
8	4.17	-6.08	17	1.25	-1.31
9	.33	-2.64	18	.77	-1.39

Table 6 (cont.)

Item	$\hat{a}$	$\hat{b}$	Item	$\hat{a}$	$\hat{b}$
			Usage		
1	1.48	-2.46	10	.78	-.83
2	.19	-10.44	11	.40	2.18
3	.74	-1.67	12	.38	5.04
4	.41	-2.52	13	.11	22.06
5	.80	-3.28	14	.04	12.49
6	.15	-1.67	15	.99	-1.11
7	.66	.58	16	.36	3.36
8	1.16	-.16	17	.50	.08
9	.50	.49	18	.32	-
			Paragraph Development		
1	1.17	-2.38	7	.62	-2.27
2	1.10	-1.42	8	.88	-1.78
3	.39	-.58	9	.55	-1.23
4	.73	2.18	10	.69	-.95
5	.29	-7.45	11	1.01	-1.68
6	.32	-2.53	12	.40	-4.15

FIGURE 1. INFORMATION FUNCTIONS FOR THE ANALYTIC RATER SPELLING RATINGS

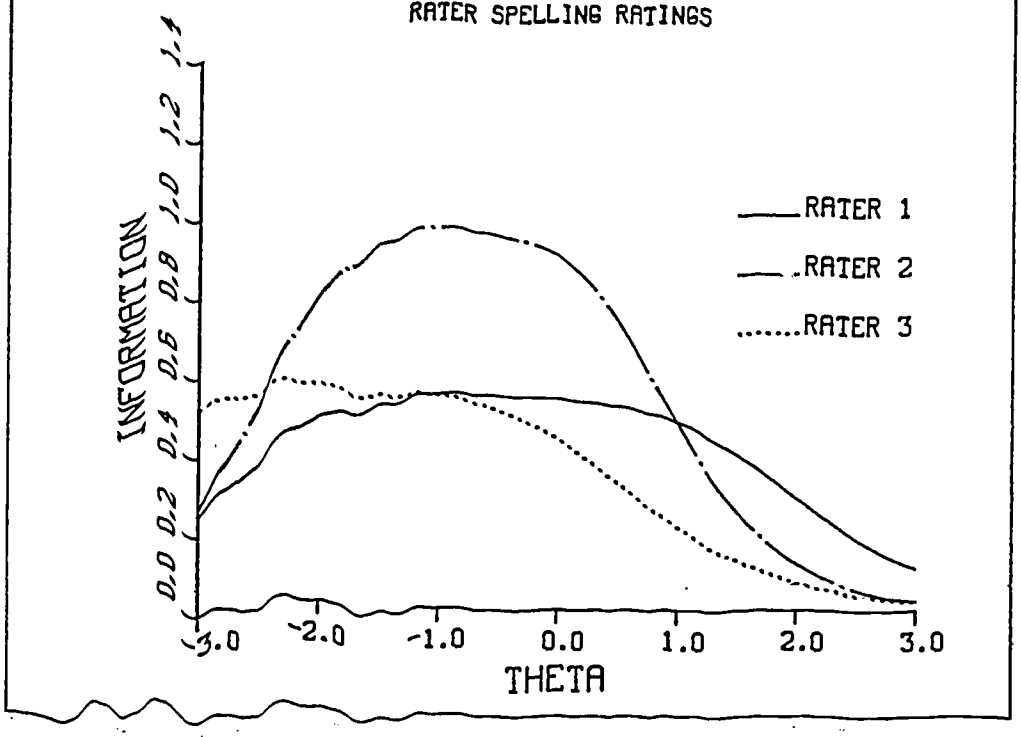

FIGURE 2. INFORMATION FUNCTIONS FOR THE ANALYTIC RATER CAPITALIZATION/PUNCTUATION RATINGS

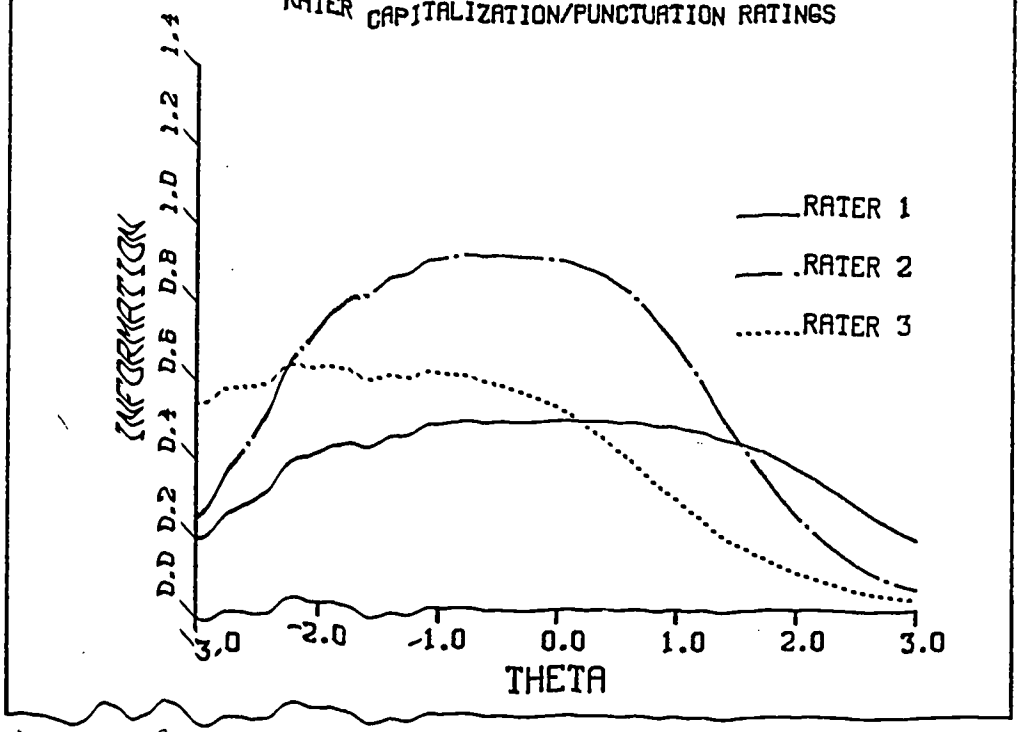

FIGURE 3. INFORMATION FUNCTIONS FOR THE ANALYTIC RATER CORRECT EXPRESSION RATINGS

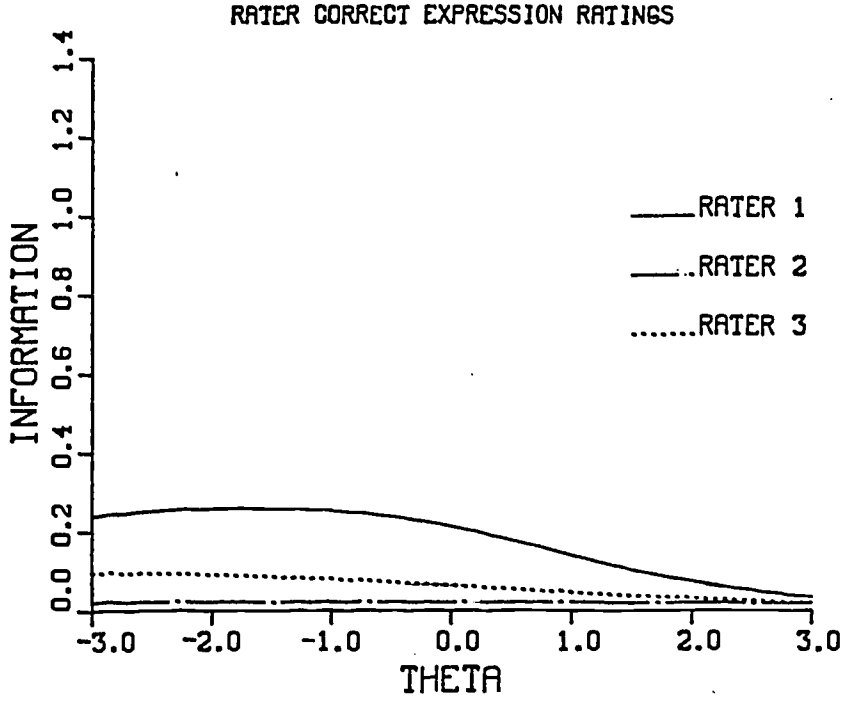

FIGURE 4. INFORMATION FUNCTIONS FOR THE ANALYTIC RATER USAGE RATINGS

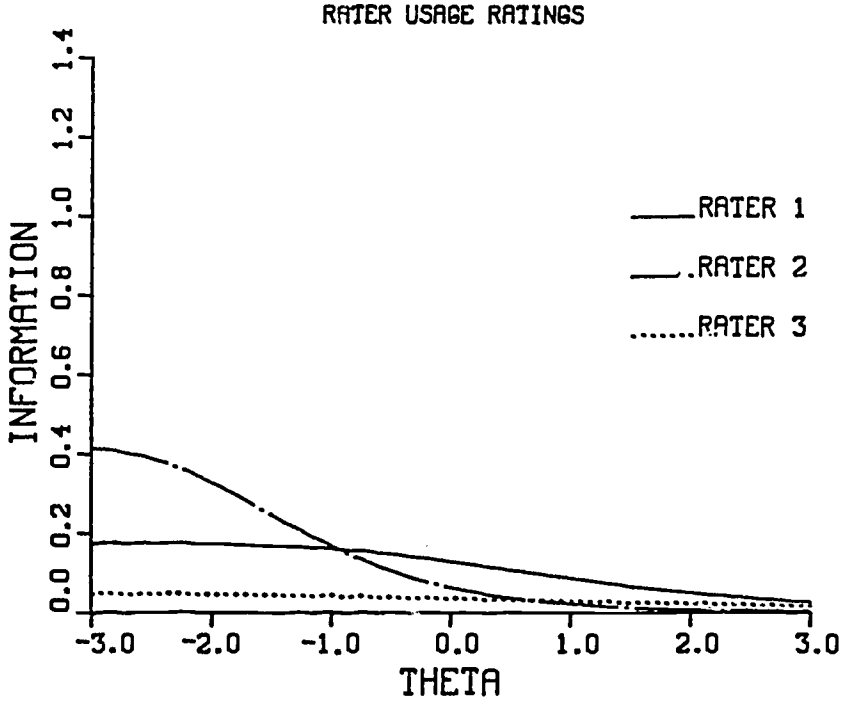

FIGURE 5. INFORMATION FUNCTIONS FOR THE ANALYTIC RATER PARAGRAPH DEVELOPMENT RATINGS

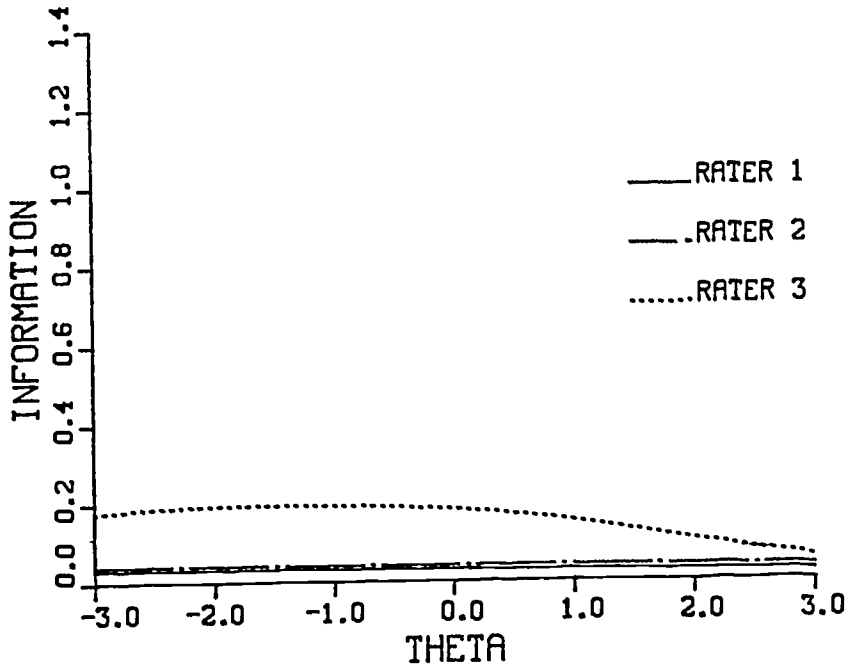

FIGURE 6. INFORMATION FUNCTIONS FOR THE TEN STANDARDIZED SPELLING QUESTIONS

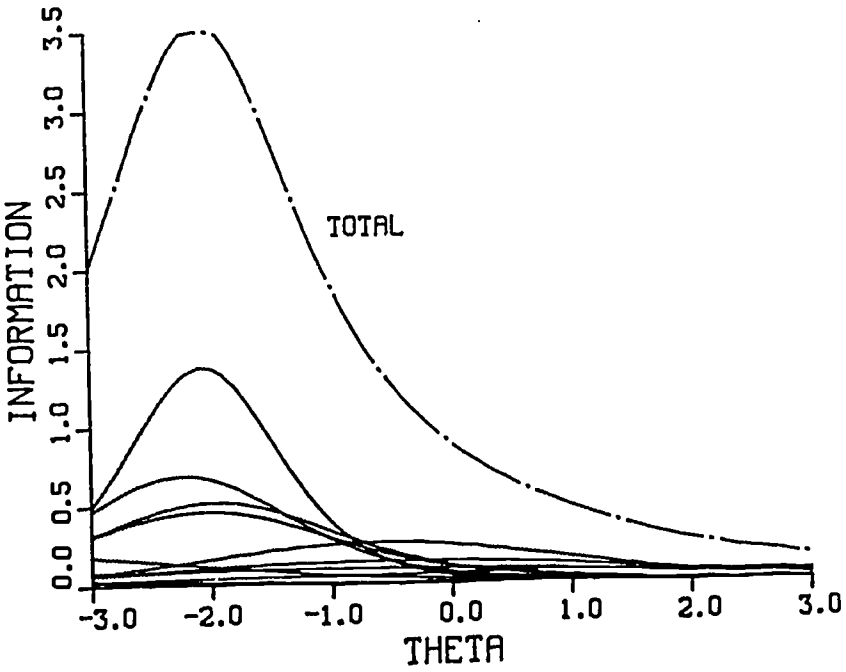

FIGURE 7. INFORMATION FUNCTIONS FOR THE TWELVE STANDARDIZED CAPITALIZATION/PUNCTUATION QUESTIONS

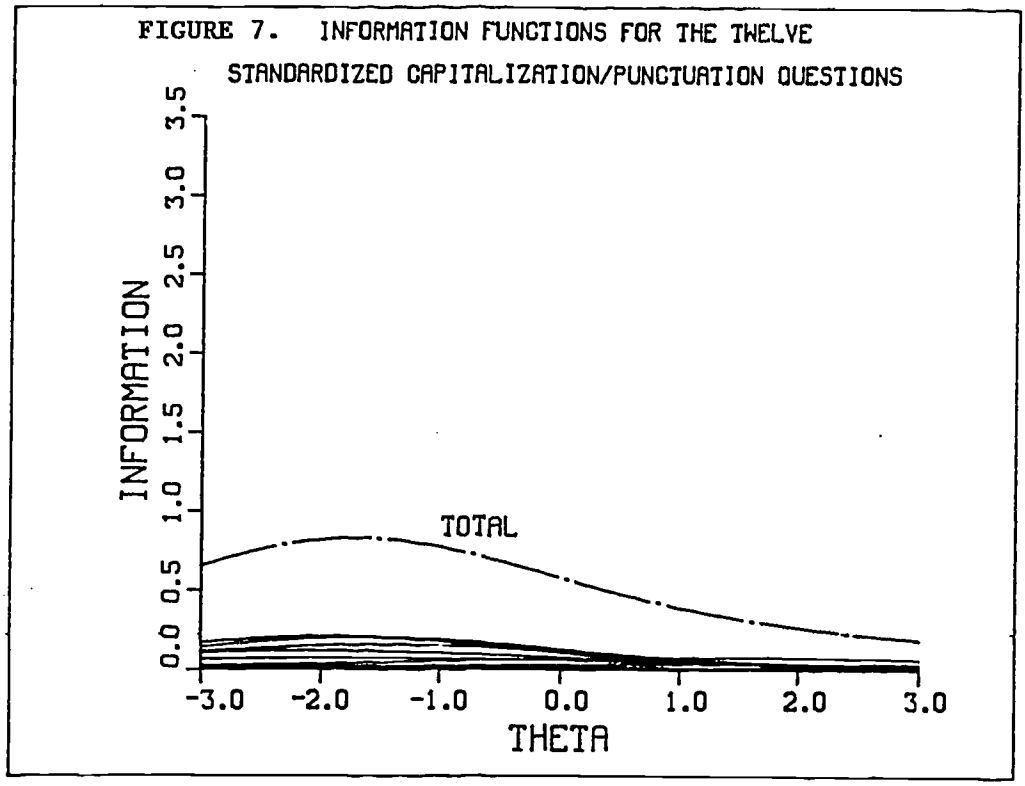

FIGURE 8. INFORMATION FUNCTIONS FOR THE EIGHTEEN STANDARDIZED CORRECT EXPRESSION QUESTIONS

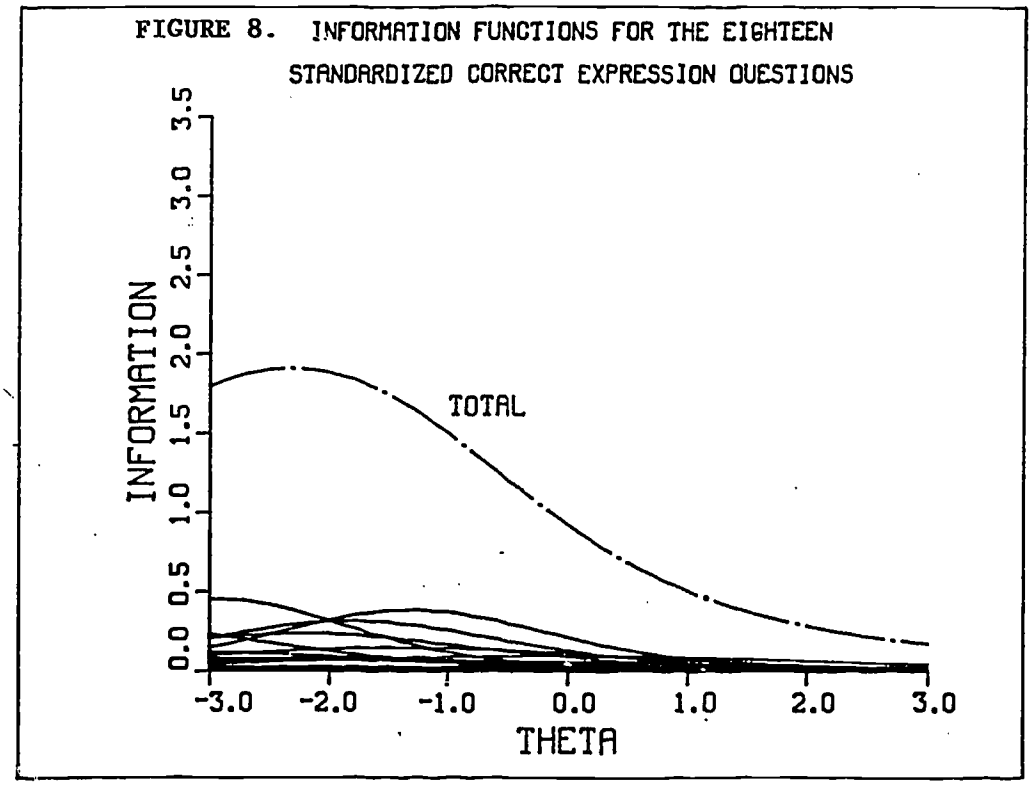

FIGURE 9. INFORMATION FUNCTIONS FOR THE EIGHTEEN STANDARDIZED USAGE QUESTIONS

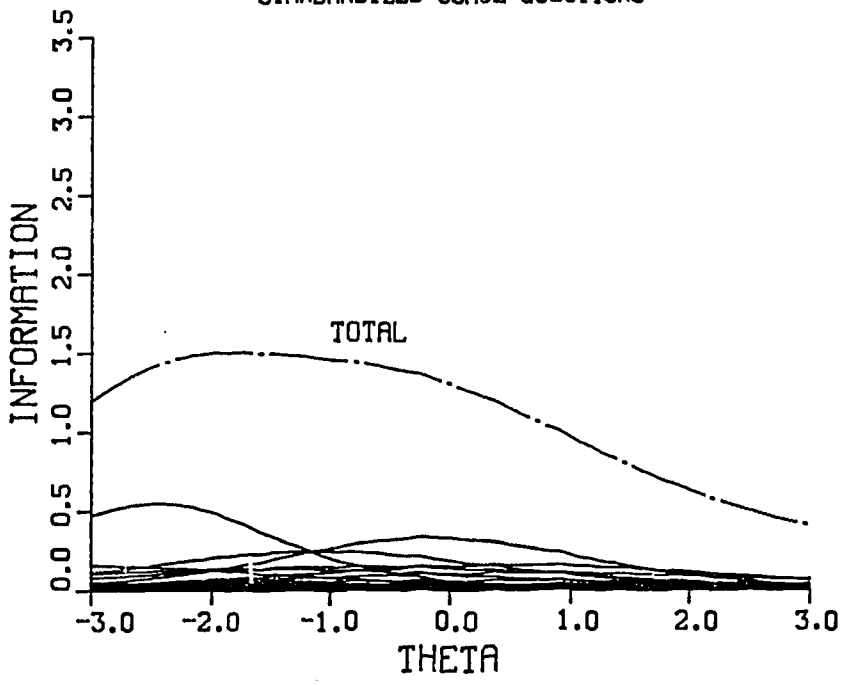

FIGURE 10. INFORMATION FUNCTIONS FOR THE TWELVE STANDARDIZED PARAGRAPH DEVELOPMENT QUESTIONS

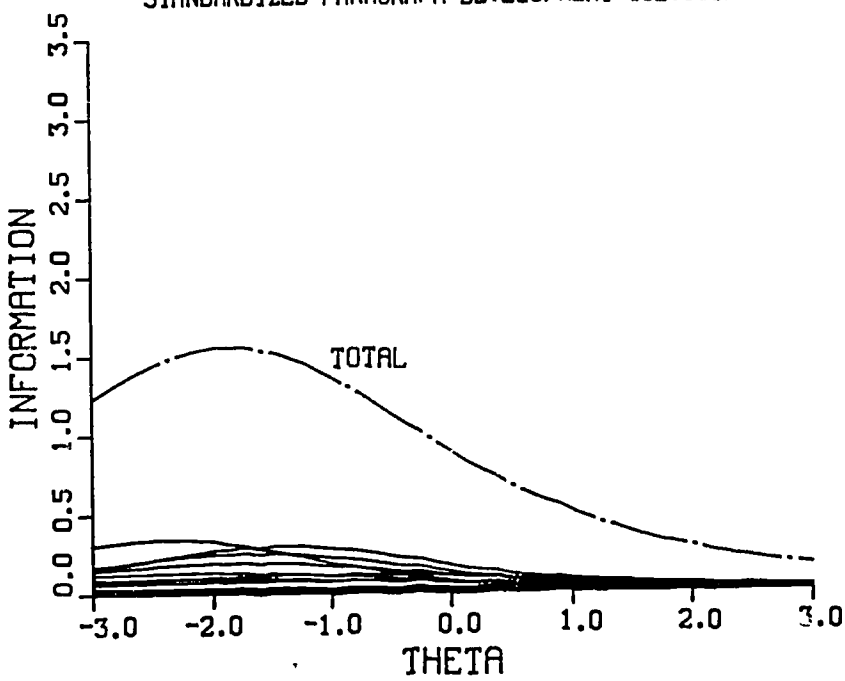

FIGURE 11. INFORMATION FUNCTIONS FOR THE AVERAGE ESSAY SPELLING RATINGS

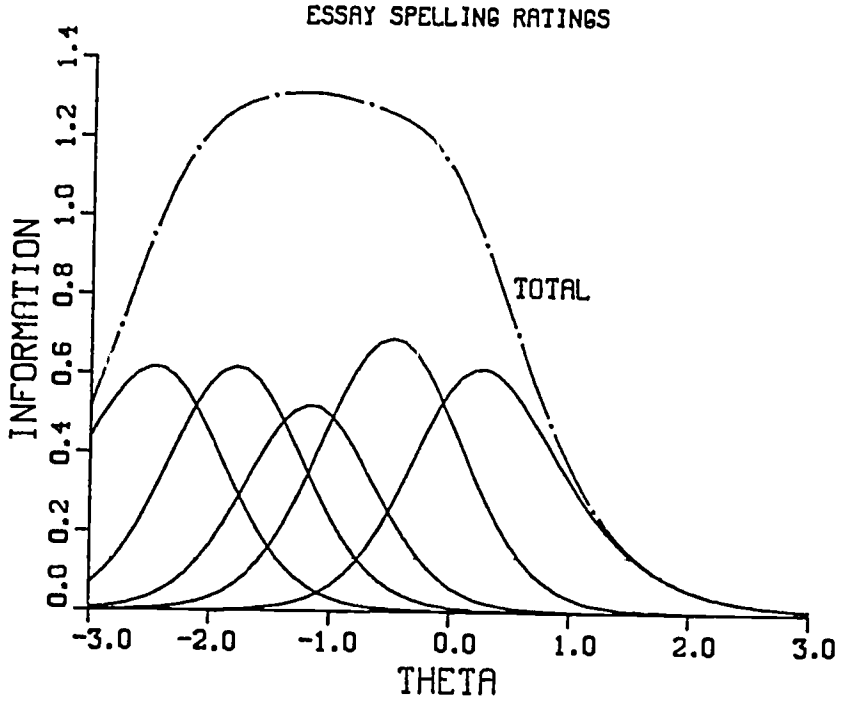

FIGURE 12. INFORMATION FUNCTIONS FOR THE AVERAGE ESSAY CAPITALIZATION/PUNCTUATION RATINGS

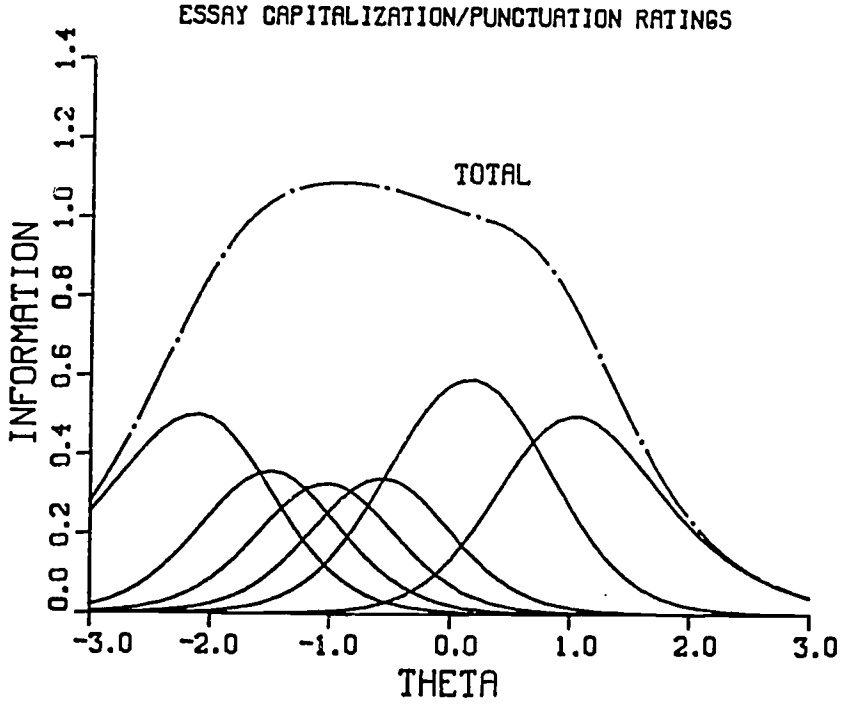

FIGURE 13. INFORMATION FUNCTIONS FOR THE AVERAGE ESSAY CORRECT EXPRESSION RATINGS

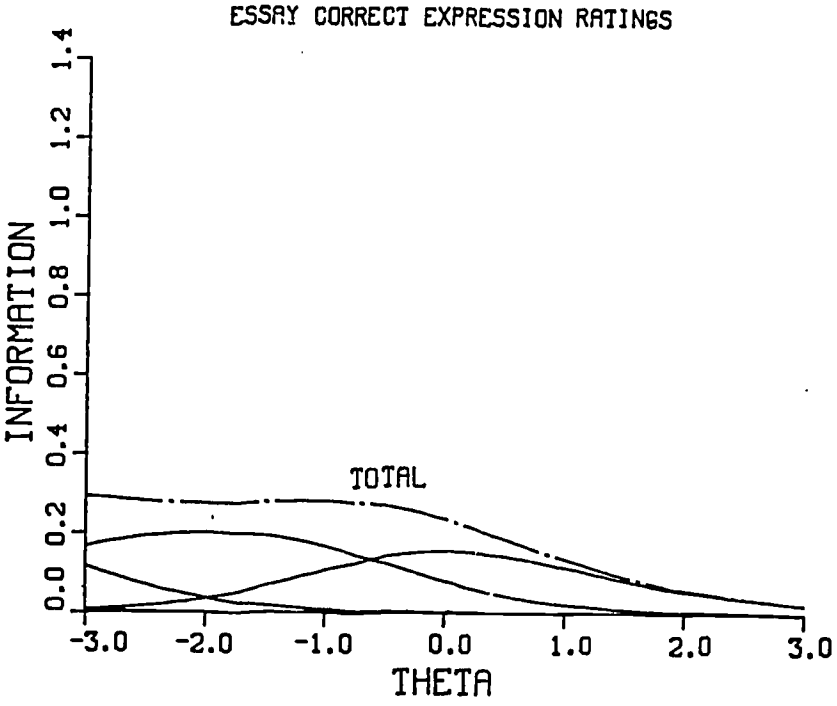

FIGURE 14. INFORMATION FUNCTIONS FOR THE AVERAGE ESSAY USAGE RATINGS

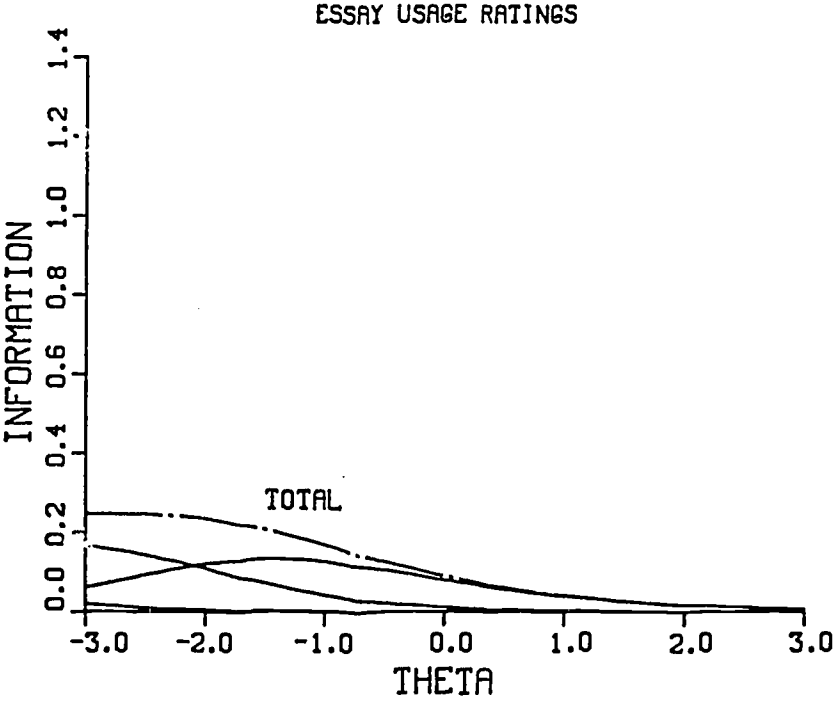

FIGURE 15. INFORMATION FUNCTIONS FOR THE AVERAGE ESSAY PARAGRAPH DEVELOPMENT RATINGS

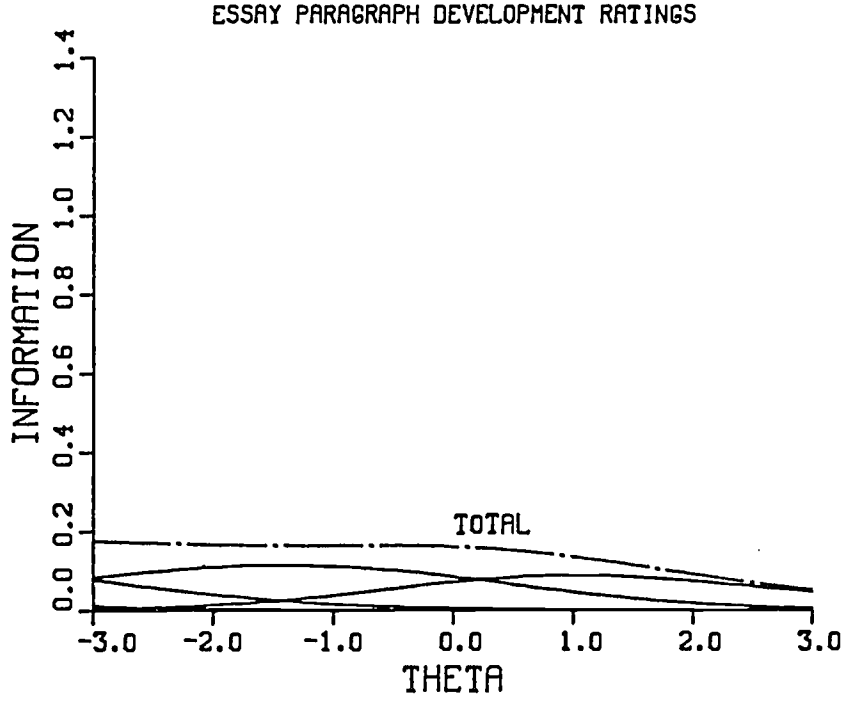

FIGURE 16. INFORMATION FUNCTION FOR THE AVERAGE ESSAY RATINGS AND STANDARDIZED TEST

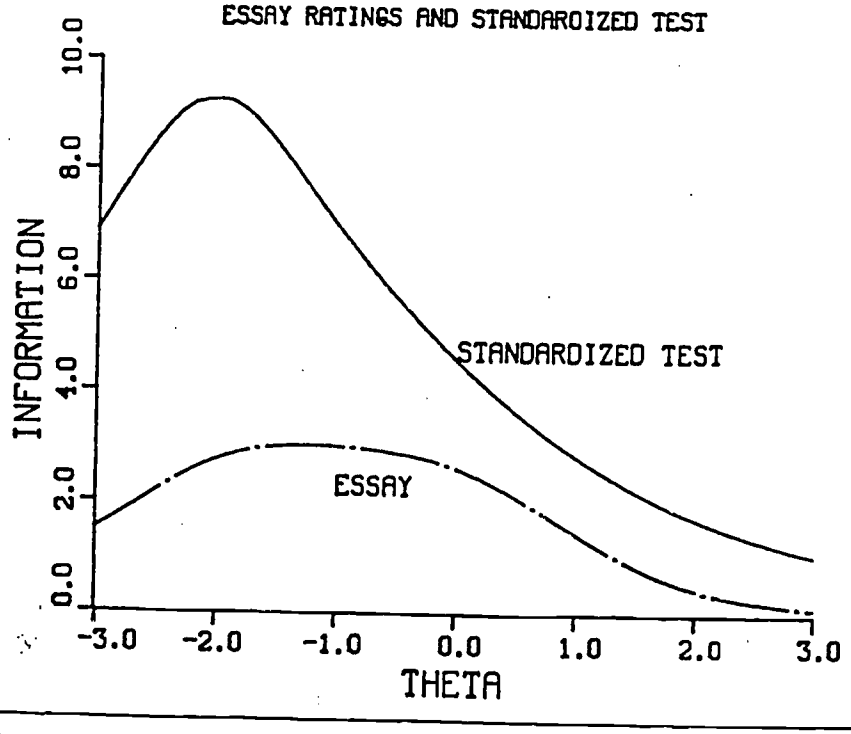

FIGURE 17. RELATIVE EFFICIENCY OF THE ESSAY TOTAL COMPARED TO THE STANDARDIZED TEST

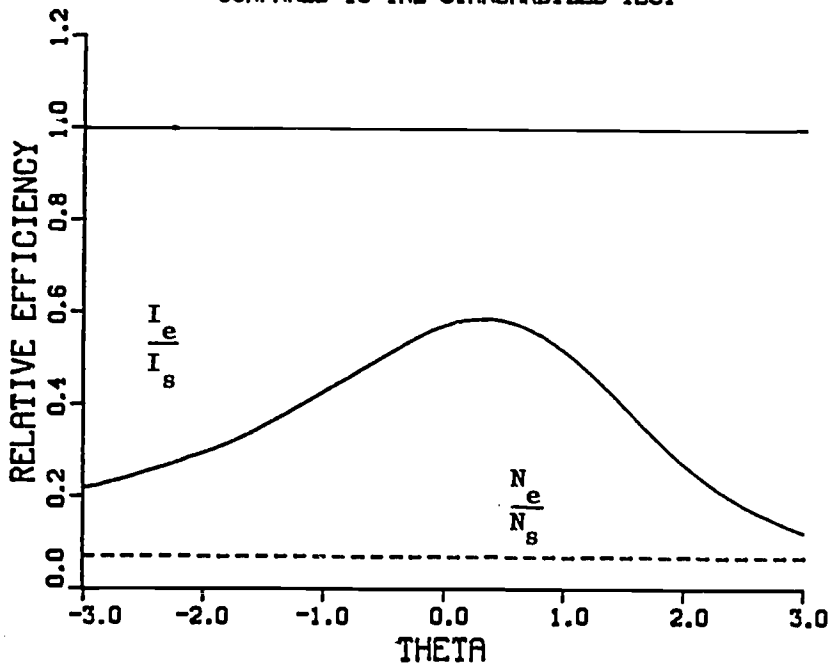

FIGURE 18. RELATIVE EFFICIENCY OF EACH ESSAY SKILL AREA RATINGS COMPARED TO THEIR STANDARDIZED COUNTERPART